SPOKES: an End-to-End Simulation Facility for Spectroscopic Cosmological Surveys

B. Nord[a], A. Amara[b], A. Bauer[h], M. Busha[e], O. Coles[j], C. Cunha[c,f], H. T. Diehl[a], J. E. Forero-Romero[g], L. Gamper[b], L. Gamper[b], B. Hambrecht[b], S. Jouvel[h], D. Kirk[j], R. Kron[a], A. Nicola[b], A. Refregier[b], W. Saunders[i], S. Serrano[h]

[a]*Fermilab Center for Particle Astrophysics, Fermi National Accelerator Laboratory, Batavia, IL 60510-0500*
[b]*ETH Zurich, Department of Physics, Wolfgang-Pauli-Strasse 27, 8093 Zurich, Switzerland*
[c]*Department of Physics, Stanford University, Stanford, CA 94305*
[d]*SLAC National Accelerator Laboratory, 2575 Sand Hill Rd., MS 29, Menlo Park, CA 94025*
[e]*Institute for Theoretical Physics, University of Zurich, 8057 Zurich, Switzerland*
[f]*Kavli Institute for Particle Astrophysics and Cosmology, 452 Lomita Mall, Stanford University, Stanford, CA, 94305*
[g]*Departamento de Física, Universidad de los Andes, Cra. 1 No. 18A-10, Edificio Ip, Bogotá, Colombia*
[h]*Institut de Ciències de l'Espai, IEEC-CSIC, Campus UAB, Facultat de Ciències, Torre C5 par-2, Barcelona 08193, Spain*
[i]*Australian Astronomical Observatory, PO Box 915 North Ryde NSW 1670, Australia*
[j]*UCL, Department of Physics & Astronomy, University College London, Gower Street, London, WC1E 6BT, UK.*

Abstract

The nature of Dark Matter, Dark Energy and large scale Gravity pose some of the most pressing questions in cosmology today. A number of wide-field spectroscopic survey instruments are being designed to meet the requirement of the high precision needed to address these fundamental questions. A key component to achieve this is the development of a simulation tool to derive the expected science performance of a given instrumental configuration, itself derived from science requirements and programmatic constraints. We describe SPOKES (**SP**ectr**O**scopic **KE**n **S**imulation), an end-to-end simulation facility for spectroscopic cosmological surveys designed to address this challenge. We use the DESpec (Dark Energy Spectrometer) experiment concept as a baseline for development, but the framework is completely general. We describe the innovative architecture of the SPOKES facility which is based on an integrated architecture, with coherent data handling and modular function access provides exact reproducibility and enables ease of use and the flexibility to evolve functions within the pipeline. The full-circle nature of the pipeline newly offers the possibility to make the science output an efficient arbiter of design optimization and feasibility testing. We present first science and performance results of the simulation pipeline. We discuss how SPOKES will provide a rigorous process to optimise the survey instrument, to demonstrate the feasibility of the measurement and to prepare for the science interpretation and exploitation of the data.

Keywords: computation, cosmology, simulation, spectroscopy, extragalactic

1. Introduction

Recent progress in cosmology in the last few decades have led to to some of the most pressing questions in fundamental science today. These are related to the understanding of dark matter, dark energy and gravity on cosmological scales. To address these questions, several wide-field spectroscopic surveys are ongoing or in the planning WiggleZ, HETDEX, SuMIRe, BigBoss, DESpec, 4MOST (Abdalla et al., 2012; Adams et al., 2010; de Jong et al., 2012; Drinkwater et al., 2010; Schlegel et al., 2011; Vives et al., 2012) . These will provide three-dimensional maps of the large-scale structure of the universe via the measurement of the angular positions and red-

Email address: nord@fnal.gov (B. Nord)

shifts of galaxies in large cosmological volumes.

The design of modern survey instruments are largely driven by the requirement of the high precision needed to address these fundamental questions. To reach this required precision, a key component is the development of a simulation tool to derive the expected science performance of a given instrumental configuration, itself derived from science requirements and programmatic constraints. This will provide a rigorous process to optimise the survey instrument, to demonstrate the feasibility of the measurement and to prepare for the science interpretation and exploitation of the data.

[We should have a few sentences regarding what simulations are being performed, but it requires a bit of care. I'm thinking just a census, not an assessment of them AR: I agree.]

In this paper, we describe SPOKES (**SP**ectr**O**scopic **KE**n **S**imulation), an end-to-end simulation facility for spectroscopic cosmological surveys. We use the DESpec (Dark Energy Spectrometer) experiment design as a baseline for development, but the framework is completely general. In §**??**, we describe the challenges which these surveys need to meet to reach the required precision, the key elements of a spectroscopic survey and the principal ingredients in a framework that simulates surveys. In §3, we present SPOKES and show how it was designed to address these challenges. In §4, we present science and performance results of the simulation pipeline. Our conclusions are summarized in §5. Details of the analysis of each function, the computing environment and the input cosmological simulation are described in the Appendix.

2. The Challenge for Spectroscopic Surveys

[AA: to Brian. The text below is very rough and no doubt full of typos, but you should be able to get the gist of what I was trying to do.]

Given the significant resources that will be invested in spectroscopic surveys, it is important to consider the new challenges that we will face at the new levels of precision that we hope to achieve. in particular, we focus here in issues associated with (i) high-precision, (ii) systematics, (iii) complexity, (iv) pre-decisions and (v) heritage.

- *High precision:* The next generation of spectroscopic surveys, as with many Stage IV (**ref**) experiments, are targeting precisions that would lead to percent level errors on the dark energy of state (w) (**ref**). These are ambitious targets that push these surveys to maximise their statistical information content by covering significant fractions of the observable sky. This means that surveys are reaching the limit of cosmic variance and so future optimisations need to be subtle and extract the maximal available information.

- *Systematics limited:* Numerous sources of systematic errors become significant as the statistical power of the surveys increases. These include ****

- *Complexities:* The difficult in dealing with systematics errors is further compounded by the fact that errors can strongly couple to each other. For instance, as an example, errors in *** can compound probes associated with *** because. Another subtlety is that the correction functions from galaxy surveys are constructed from the positions of samples (in this case the galaxies) rather a correlation function of the tags (e.g. size, ellipticity etc). This means that subtle systematics that effect the spacial selection function can have leading order effects, rather than second order in the case of tag correlations, unless they are accounted for carefully.

- *Pre-decisions and target selection:* Spectroscopic surveys differ crucially from imaging surveys in that key decisions about the target sample need to made beforehand. This complexities the possible instrument configurations increases the importance of modelling the system early before data is collected.

- *Heritage work/code:* Wide-field spectroscopic surveys for large scale structure mapping is a matured field. This means that there are many tools and methods that have been developed to

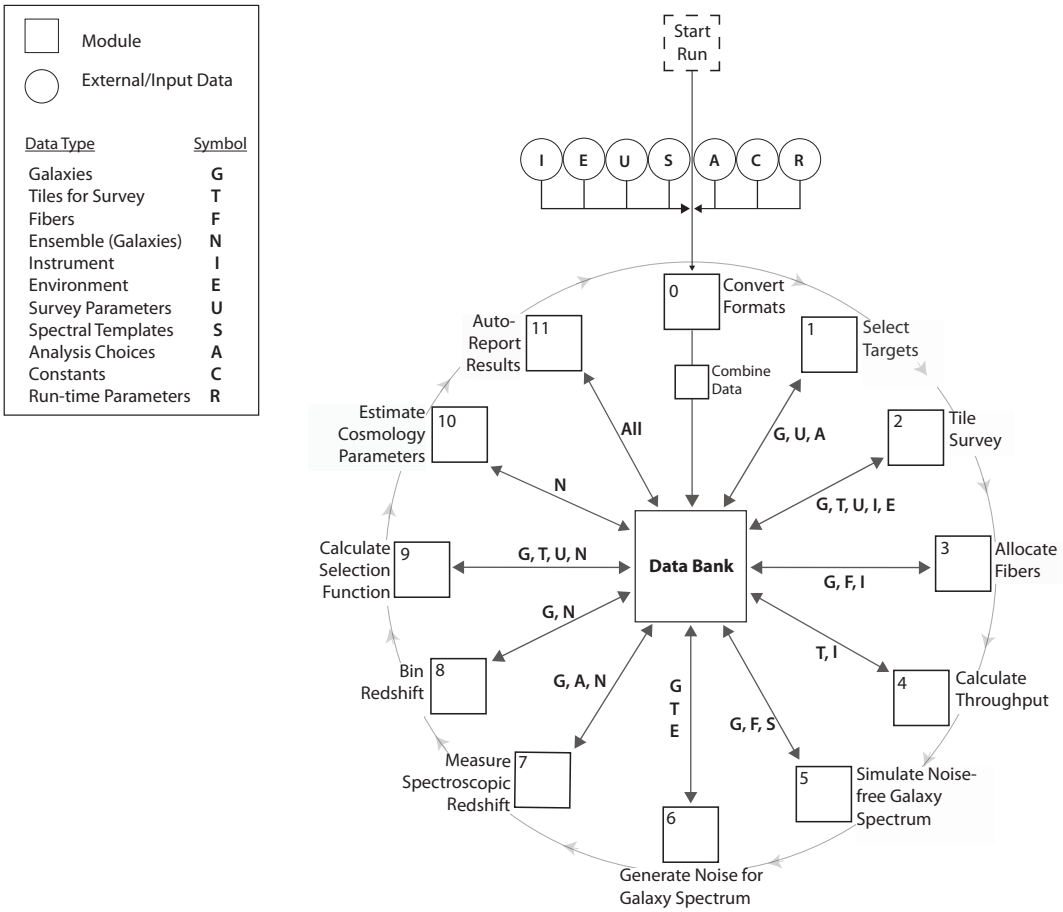

Figure 1: The SPOKES pipeline flow diagram shows the sequence of calculations viz. the main modules (labeled 0-11 in the upper left corner of the boxes), as well as data exchange between modules via the data bank. All external data (circles) are ingested before the pipeline is run and placed into the data bank for access downstream: input data is converted into the SPOKES native data format, and all data and parameters are agglomerated into one catalog. The modules (squares) access (arrows) the data from the data bank via the API send their output back into the data bank. The clock-like nature of the pipeline emphasizes this data management scheme as well as the action sequence.

date. Incorporating such expertise is therefore highly desirable, but the difficulty is that these available tools form a strongly heterogeneous sets making the process of integrating them difficult.

It is becoming increasingly clear that because of these difficulties we will be ever more reliant on simulations to design and verify future experiments. Further more, as the simulations become more sophisticated they are also likely to play more important roles in developing the data processing framework. To full-fill their potential, simulation pipelines for spectroscopic surveys require several key ingredients. Because of the potential for complex coupling of systematics errors the simulations will need to track all the important steps in the process. This points towards the need for an *end-to-end* architecture.

[AA: stuff below still needs some polish]

All the functions should pass (meta)data consistently and clearly from one function to the next. The architecture will be sufficiently *integrated* such that each function communicates with the rest of the pipeline through the same mechanisms to allow data and logic to be tracked precisely. This will additionally allow *reproducibility* at a very fine level: one should be able to produce identical results with identical inputs; e.g., stochasticity can be removed/controlled by saving the value of random seeds.

The framework will still, however, be sufficiently *flexible* to permit ingestion of new functions and modification of current functions. In addition, some operations of the experiment or analysis of the data will be time-consuming or memory-heavy.

Therefore, the pipeline will adapt across a range of run modes—at one side to accomodate high-speed, high-efficiency runs and at the other to permit computationally-intense (e.g., image-level) calculations, which may require parallelization: the pipeline will be sufficiently flexible to accomodate a *high dynamic range* in detail and parallelization.

The run time of the pipeline will depend on the run mode (i.e., the level of detail simulated). However, there will be a mode that can both run *fast* and relatively precisely to recover the correct output of the experiment.

3. SPOKES Pipeline

We ingest the pipeline functions into a simulation architecture designed to meet the requirements. Fig 1 shows the arrangement of components along with the architectural components that we describe below—from the data format, to the module functions to the glue that holds it all together.

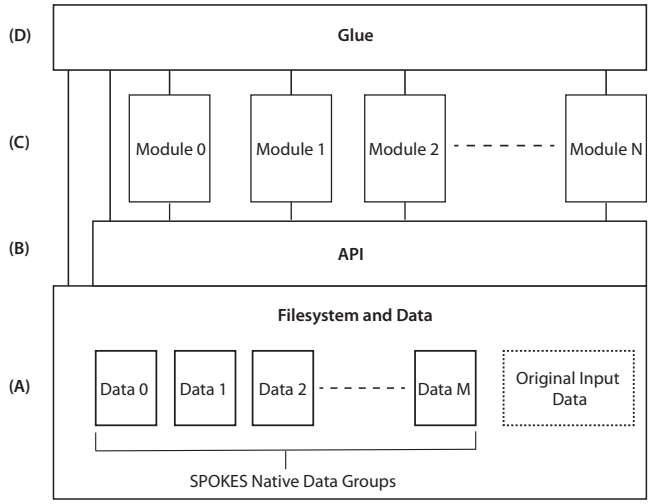

Figure 2: Data in the data bank (Section A) is handled by three principal elements. The API (B) handles data access throughout the pipeline, starting with conversion of data to the SPOKES native data format. The modules (C) read and write data via the API, but they are otherwise independent of one another. The glue (D) connects the various elements of the pipeline, uses the API to access data, and is the main interface for the user—i.e., the point where the user selects the order of modules and the versions to use.

3.1. Glue

Finally, the topmost layer, the **glue** (Fig. 2D), coheres all of the pieces into one point of configuration. The glue is responsible for merging configuration files, combining data and parameters into the data bank, managing the modules (which are executed and in what order), and finally the execution of the pipeline.

3.2. Modules

Pipeline operations are broken into discrete *modules*, each of which performs a specific function in the pipeline. Each of these only handles specific data sets. The modules are functionally independent of one another, but linked via the base (data) layer of the architecture (see Fig. 2C). The modules are discretized in such a way as to allow for one of them to be replaced by a new algorithm, not affecting the remainder of the modules, as long as the same data are passed back into the data layer. The principal constraint on module ingestion is the data available in the databank: new implementations of modules are permitted given that they conform to the structure and availability of data. While the architecture is modular, we integrate the subsystems into the pipeline to ensure that the pipeline runs flawlessly from end to end.

We provide a brief, high-level overview of how a spectroscopic survey is planned and undertaken. We delineate sequentially the operations and science analysis—from creating/gathering the input to redshift analysis to cosmological parameter estimation.

1. An imaging survey (e.g., SDSS, DES, LSST) finds galaxies in the sky and measures their photometry (magnitudes and colors) in a set of wide-band filters. The photometry and sky position is then used to decide if a galaxy contains a spectrum of interest.
2. In similar fashion to an imaging experiment, the spectroscopic experiment surveys the sky tile by tile, successively.
3. For each tile, the fibers are allocated to targeted galaxies.
4. A galaxy's spectrum is then measured in the spectrograph or reconstructed in simulations. In simulation, the spectra are reconstructed from photometrically-derived coefficients (from the imaging sample) in tandem with empirically-derived spectral templates of known galaxy/spectrum types.

5. In simulation, we replicate the instrument optical throughput and noise sources (i.e., atmosphere, CCD read noise), which obstruct and confuse photon paths, reducing the overall signal-to-noise of mock observed spectra.
6. By comparing the survey target spectra with that of galaxies with known redshifts, we can measure spectroscopic redshifts of the new galaxies.
7. With a number density distribution of spectroscopic redshifts, Fisher matrix analyses can provide estimates of cosmological parameters.

Each component and algorithm employed is discussed in further detail in Appendix A.

3.3. Data Handling

An application programming interface (API) provides a simple, user-friendly and robust interface between the module and the data bank (see Fig. 2B). The API is designed to efficiently handle data, simplify data access and reduce possible bugs in input and output. All the modules use the same format for reading and writing data: this standard reduces the possibility of having inconsistent or duplicate data. It makes the data readily accessible for quality assurance tests and allows the flexible deployment of different data sets.

3.4. Data Location and Format

Our data format has to be able to handle many data types, perform well while handling large amounts of data and be flexible enough to store all data for a rapidly evolving pipeline. We avoid having to maintain many input files and having to document the contents and source of each file separately. One of the main challenges was to implement an exchange scheme between modules that is easy to use, flexible and robust, so that the substitution of a module does not break the pipeline.

The FITS data format[1] is the main format for astronomical imaging and catalogs in the modern era (second perhaps only to ASCII), and has a long history (e.g. White et al., 1991). We evaluated the FITS format and found that it was not flexible enough for the requirements described above.

[1] http://fits.gsfc.nasa.gov/

We thus introduce the concept of the **databank**: in SPOKES, this data bank is a single HDF5 format (Group, 2000-2010) file containing all the data used within the pipeline. A HDF5 file is structured like the filesystem on a harddisk, where each data set has a unique path—as in */group/subgroup/dataset*, which resides in a named "group" and "subgroup." These data sets can contain nearly any data type, including arrays. The databank need not be a single HDF5 file however; the data can be partitioned in whichever way(s) that the modules or run modes require.

The chosen data format enhances readability and clarity, and it provides modularity of data access: a module may use individual aspects of a data group without having to read in all data. For example, a module can access galaxy identification numbers and positions without reading all the other data that another module might need.

All original data from input, including parameters (e.g., telescope optics choices) and data (e.g., galaxies), are converted to the native data format and separated into *M* data sets within the databank upon initiation of the pipeline, as shown in the base data layer of the architecture in Fig. 2A.

The data groups within the data bank are partitioned according to module usage and related information; these data groups are shown in the legend in Fig 1. In sum, a **data set** is contained within a data *subgroup* of a data *group*. For example of immutable parameters, "Spectrograph" is a subgroup of "Instrument," and "wavelength range" is the data set of interest. For an example of changing data, "Galaxy" is another main group, with a data sets "RA" and "spectroscopic redshift": the former doesn't change, while the latter is created in Module 7. Our data sets are organized to coincide with the data handling methodology that allows read in of specific data when they are needed, a property we call "What you need when you need it" or *WYNWYNi*. The data groups are described below:

- Galaxies (*G*) contains all galaxy data.
- Survey Tiles (*T*) contains a set of tile information (sky position, airmass, time of observation, etc) and is used to link galaxies with the time and observation environment in which they were observed.

Table 1: Input Parameters

Data Group	Used by Module(s)	Parameter	Value(s)
Instrument (*I*): Fibers			
	3,4	Fiber diameter	1.27 arsec
	3	Pitch	6 mm
	3	Patrol radius	6 mm
	3	Number of Fibers	5000
	2,3	Fiber Arrangement	Hexagon
	2,3	Passes per Tile	2
Instrument (*I*): Telescope			
	4	Optical Efficiency*	0.25
Instrument (*I*): Spectrograph			
	6	Read Noise	5 photons
	6,7	Wavelengths	[350, 1050] nm
Survey Parameters (*U*)			
	2	Exp time	1200 s
	2,9,10	Area	50 sq. deg.
	2	Duration	500 Nights
	1,2,9	RA range	$[xx, xx]$
	1,2,9	DEC range	$[xx, xx]$
Environment (*E*): Atmosphere**			
	2	Weather/Seeing Model	
	6	Sky Background	Gemini Sky Models
	6	Atmospheric Extinction	Palomar Extinction Curves
Analysis (*A*): Target Selection			
	1	$(g - r)$ color	[-1, 2]
	1	$(r - i)$	[-1, 2]
	1	i mag	< 23.5
	1,7	photo-z range	$z_{photo} < 1$
Analysis (*A*): Redshift Binning			
	10	bin width	0.1
	10	max redshift	1.0

The input variables to the pipeline and the values used in the demonstration run of the SPOKES pipeline. Data groups and module numbers coincide with those of Fig 1. The 23 parameters shown here are necessary to running a spectroscopic experiment; some may indeed by derived from more fundamental parameters, but these are required.

* This value is the mean at the plateau of the throughput spectrum. See Appendix A.4 for details of the throughput calculation.
** see Appendix A.5 for details

- Fibers (*F*) contains a set of fiber information (location in focal plane,
- Ensemble (*N*) contains data on the galaxies as a collection—the redshift histogram, related cosmological constraints, etc.
- Instrument (*I*) contains several subgroups representing the subsystems of the instrument—optics, fibers and spectrograph—each of which has several parameters.
- Environment (*E*) contains the information regarding the atmosphere (absorption and emission spectra) and location (e.g., elevation) at which the observations are taking place.
- Survey Parameters (*U*) holds the data necessary to run the survey—e.g., exposure time per tile and region of the sky to be observed.
- Spectral Templates (*S*) contain the eigentemplates used to reconstruct galaxy spectra.
- Analysis Choices (*A*) contains the information that can be used to vary the analysis methods—

e.g, magnitude or color cuts for Target Selection (Module 1) and bin size for redshift binning (Module 8).

- Constants (C) holds physical constants and the random seed.
- Run-time Parameters (R):

The formal delineation of data sets also elucidates the interaction between them. For example, galaxies are linked to the tiles in which they are observed, but it is inefficient for each galaxy to directly carry the information for each tile. Therefore, each galaxy merely carries a Tile identification number (or -1 if not observed), which is used to reference the tile catalog; the tile catalog holds all the information about each tile. In this way, the tile fiber and galaxy data sets interact as if in a relational database.

The input parameters to the pipeline represent the hardware and analysis characteristics. The parameters are different from the data in that they are fewer, immutable during a pipeline run and are mostly input by hand (with the exception of telescope optics; see Appendix A.4). We see the complete list of parameters in Table 1. Note that several modules use the same parameters.

3.5. Quality Assurance

Our Quality Assurance (QA) paradigm aims at a continuous integration, performing tests at both the unit and facility levels to constantly monitor the logical and programmatic progression of the pipeline. It provides sufficient information to diagnose outcoming science results at each step, and trace back the affects of each component on the final results.

QA first performs a series of basic logical cross-checks at the onset of each module: e.g., after fibers have been allocated to galaxies, the same function checks that each galaxy has received some value flagging it as observed or not; but, it also checks that this value is within the range of available fiber indices.

The pipeline also provides scientific diagnostic figures and plots to check fidelity at each step: e.g., the Survey Strategy produces a map of the fields observed, and the spectral reconstructor and noise generator produce images of spectra for a subsample of the mock-observed galaxies.

[describe figures]

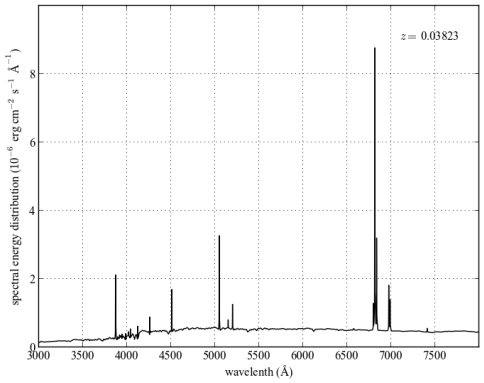

Figure 3: A pure galaxy spectrum reconstructed from a set of templates. This process is described in Appendix A.5. Note the principal features of the spectrum, as well as the redshifting.

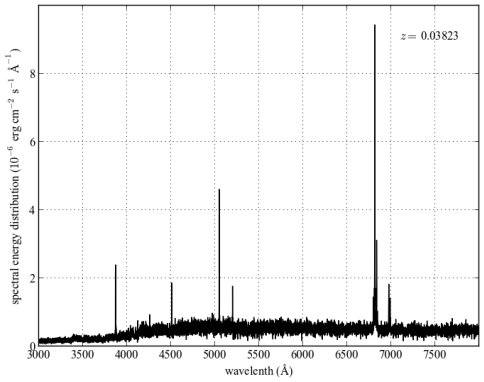

Figure 4: A galaxy spectrum with noise computed from multiple sources, including Poisson noise from photon counts and CCD readout noise. Details of the noise generation process can be found in Appendix A.5.[could overlay the pure on top of this one?]

4. Results

Here, we show the results for a run of the SPOKES pipeline with a particular set of input parameters and mock galaxies, with the intent of producing sensible cosmological measurements, compared with already-complete and analyzed surveys. First, we describe the choices for the input parameters describing the telescope, survey and analysis choices. Then, we show the science results and compare them to results derived from past and modern surveys. Finally, we detail the computational performance of the pipeline.

4.1. *Data and Parameters*

We take standard inputs (both data and parameters) and pass it through standardized modules allowing the parameter-specified modules to act on the data and produce standard outputs. This provides a clean and integrated analysis tool.

[Describe our choices of parameters.]

4.2. *Science Performance*

In order to verify the capabilities of SPOKES to predict the outcome of future surveys, we run the pipeline for a typical set of parameters for a survey, monitoring key outputs to assess to the pipeline performance. The key outputs to assess are 1) a comparison of the spectroscopic and the true redshifts of galaxies, 2) the number distribution of spectroscopic redshifts and 3) the resultant DE Figure of Merit, confidence contours in the $w_0 - w_a$ plane. As we merely seek to verify that the pipeline results are sensible to demonstrate performance, we do not here address already-complete surveys; this comparison is reserved for later work **REF(Nord, et al., 2013b)**, where we assess SPOKES output for parameters from a host of completed surveys. Table 1 shows the values chosen for our demonstration run.

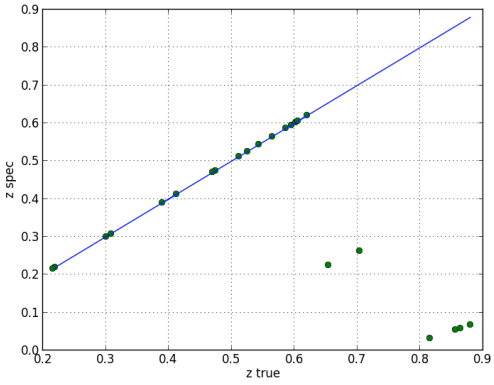

Figure 5: Comparison of *true* redshifts and the measured *spectroscopic* redshifts. See Appendix A.6

4.2.1. *More specific questions*

- Will's suggestions
- Compare to only having and exposure time calculator.

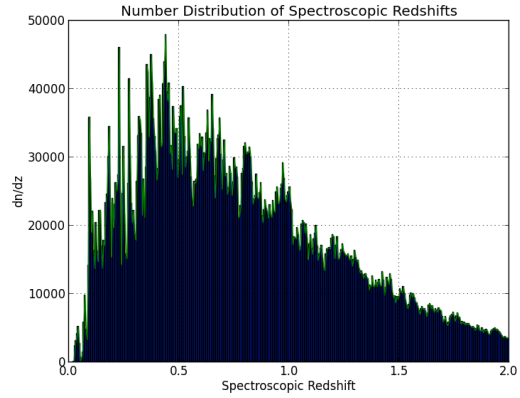

Figure 6: Redshift distribution $\mathrm{d}n/\mathrm{d}z\,(z)$ for the spectroscopic redshifts (bars) and for the true galaxy redshift (line). The bin width is set 0.1, producing 19 bins between z=0 and z=2. See Appendix A.7

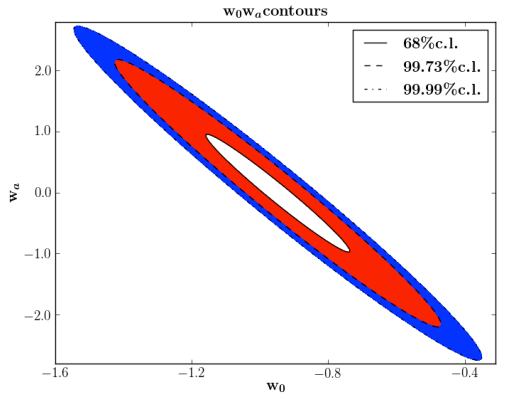

Figure 7: Confidence contours for a joint estimate of dark energy parameters, w_0 and w_a. See Appendix A.9

4.3. *Computational Performance*

For the simulation demonstration runs, we report on computational performance to show that the SPOKES pipeline has the key ingredients outlined in §**??**. We show here the timing and memory usage for one of our typical runs.

5. Final Remarks and Conclusions

Modern cosmology experiments have become sufficiently precise and complex such that new methods are required to perform accurate feasibility studies and to perform survey optimization. We have demonstrated the completeness, speed and flexibility of the SPOKES simulation pipeline.

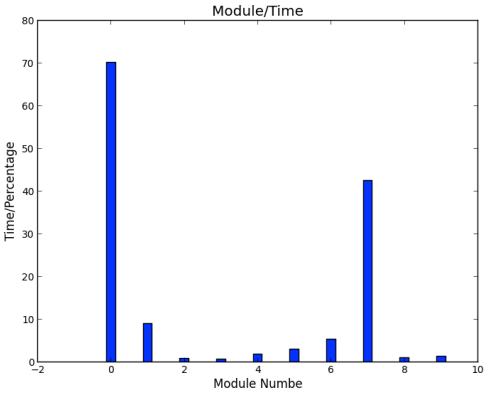

Figure 8: The timing and memory suage of each module.

The authors would like to thank many people for useful discussions during development of the pipeline and in the writing of this manuscript, including Michael Meyer at ETH.

References

Abdalla, F., Annis, J., Bacon, D., Bridle, S., Castander, F., Colless, M., DePoy, D., Diehl, H. T., Eriksen, M., Flaugher, B., 2012. The Dark Energy Spectrometer (DESpec): A Multi-Fiber Spectroscopic Upgrade of the Dark Energy Camera and Survey for the Blanco Telescope. arXiv preprint arXiv:1209.2451.
URL http://arxiv.org/abs/1209.2451

Adams, J. J., Blanc, G. A., Hill, G. J., Gebhardt, K., Drory, N., Hao, L., Bender, R., Byun, J., Ciardullo, R., Cornell, M. E., Finkelstein, S. L., Fry, A., Gawiser, E., Gronwall, C., Hopp, U., Jeong, D., Kelz, A., Kelzenberg, R., Komatsu, E., MacQueen, P. J., Murphy, J., Odoms, P. S., Roth, M., Schneider, D. P., Tufts, J. R., Wilkinson, C. P., Dec. 2010. THE HETDEX PILOT SURVEY. I. SURVEY DESIGN, PERFORMANCE, AND CATALOG OF EMISSION-LINE GALAXIES. The Astrophysical Journal Supplement Series 192 (1), 5.
URL http://stacks.iop.org/0067-0049/192/i=1/a=5?key=crossref.f41345950d93ab1bb4e5a72fdbfa2bc9

Blanton, M. R., Lin, H., Lupton, R. H., Maley, F. M., Young, N., Zehavi, I., Loveday, J., Apr. 2003. An Efficient Targeting Strategy for Multiobject Spectrograph Surveys: the Sloan Digital Sky Survey "Tiling" Algorithm. The Astronomical Journal 125 (4), 2276–2286.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2003AJ....125.2276B&link_type=ABSTRACT

Busha et al., 2013. Addgals ii. arXiv.org.

Conroy, C., Wechsler, R. H., KRAVTSOV, A. V., Mar. 2007. The Hierarchical Build-Up of Massive Galaxies and the Intracluster Light since z=1. arXiv.org astro-ph.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2007ApJ...668..826C&link_type=ABSTRACT

Cunha, C. E., Huterer, D., Lin, H., Busha, M. T., Wechsler, R. H., Jul. 2012. Spectroscopic failures in photometric redshift calibration: cosmological biases and survey requirements. arXiv.org 1207, 3347.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2012arXiv1207.3347C&link_type=ABSTRACT

de Jong, R. S., Bellido-Tirado, O., Chiappini, C., Depagne, É., Haynes, R., Johl, D., Schnurr, O., Schwope, A., Walcher, J., Dionies, F., Haynes, D., Kelz, A., Kitaura, F. S., Lamer, G., Minchev, I., Müller, V., Nuza, S. E., Olaya, J.-C., Piffl, T., Popow, E., Steinmetz, M., Williams, M., Winkler, R., Wisotzki, L., Ansorgb, W. R., Banerji, M., Solares, E. G., Irwin, M., Kennicutt, Jr, R. C., King, D., McMahon, R., Koposov, S., Parry, I. R., Walton, N. A., Finger, G., Iwert, O., Krumpe, M., Lizon, J.-L., Vincenzo, M., Amans, J.-P., Bonifacio, P., Cohen, M., Francois, P., Jagourel, P., Mignot, S. B., Royer, F., Sartoretti, P., Bender, R., Grupp, F., Hess, H.-J., Lang-Bardl, F., Muschielok, B., Böhringer, H., Boller, T., Bongiorno, A., Brusa, M., Dwelly, T., Merloni, A., Nandra, K., Salvato, M., Pragt, J. H., Navarro, R., Gerlofsma, G., Roelfsema, R., Dalton, G. B., Middleton, K. F., Tosh, I. A., Boeche, C., Caffau, E., Christlieb, N., Grebel, E. K., Hansen, C., Koch, A., Ludwig, H.-G., Quirrenbach, A., Sbordone, L., Seifert, W., Thimm, G., Trifonov, T., Helmi, A., Trager, S. C., Feltzing, S., Korn, A., Boland, W., Jun. 2012. 4MOST - 4-metre Multi-Object Spectroscopic Telescope. arXiv.org astro-ph.IM.
URL http://arxiv.org/abs/1206.6885v1

Drinkwater, M. J., Jurek, R. J., Blake, C., Woods, D., Pimbblet, K. A., Glazebrook, K., Sharp, R., Pracy, M. B., Brough, S., Colless, M., Couch, W. J., Croom, S. M., Davis, T. M., Forbes, D., Forster, K., Gilbank, D. G., Gladders, M., Jelliffe, B., Jones, N., Li, I.-h., Madore, B., Martin, D. C., Poole, G. B., Small, T., Wisnioski, E., Wyder, T., Yee, H. K. C., Jan. 2010. The WiggleZ Dark Energy Survey: survey design and first data release. Monthly Notices of the Royal Astronomical Society 401 (3), 1429–1452.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2010MNRAS.401.1429D&link_type=ABSTRACT

Group, T. H., 2000-2010. Hierarchical data format version 5. arXiv.org, 1–11.
URL http://www.hdfgroup.org/HDF5

Hu, W., Jain, B., Aug. 2004. Joint galaxy-lensing observables and the dark energy. Physical Review D 70 (4), 43009.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2004PhRvD..70d3009H&link_type=ABSTRACT

Kendall, G., Stuart, A., 1973. The Advanced Theory of Statistics. Vol.2: Inference and: Relationsship. Griffin.
URL http://books.google.ch/books?id=elabQwAACAAJ

KRAVTSOV, A. V., Berlind, A. A., Wechsler, R. H., KLYPIN, A. A., Gottloeber, S., Allgood, B., Primack, J. R., Aug. 2003. The Dark Side of the Halo Occupation Distribution. arXiv.org astro-ph.
URL http://arxiv.org/abs/astro-ph/0308519v2

Limber, D. N., 1953. The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field. The Astrophysical Journal.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1953ApJ...117..134L&link_type=ABSTRACT

Linder, E. V., Mar. 2003. Exploring the Expansion History of the Universe. Physical Review Letters 90 (9), 91301.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2003PhRvL..90i1301L&link_type=ABSTRACT

Loveday, J., Norberg, P., Baldry, I. K., Driver, S. P., Hopkins, A. M., Peacock, J. A., Bamford, S. P., Liske, J., Bland-Hawthorn, J., Brough, S., Brown, M. J. I., Cameron, E., Conselice, C. J., Croom, S. M., Frenk, C. S., Gunawardhana, M., Hill, D. T., Jones, D. H., Kelvin, L. S., Kuijken, K., Nichol, R. C., Parkinson, H. R., Phillipps, S., Pimbblet, K. A., Popescu, C. C., Prescott, M., Robotham, A. S. G., Sharp, R. G., Sutherland, W. J., Taylor, E. N., Thomas, D., Tuffs, R. J., van Kampen, E., Wijesinghe, D., Feb. 2012. Galaxy and Mass Assembly (GAMA): ugriz galaxy luminosity functions. Monthly Notices of the Royal Astronomical Society 420 (2), 1239–1262.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2012MNRAS.420.1239L&link_type=ABSTRACT

McBride, C., Berlind, A., Scoccimarro, R., Wechsler, R., Busha, M., Gardner, J., van den Bosch, F., Jan. 2009. LasDamas Mock Galaxy Catalogs for SDSS. American Astronomical Society 213, 253.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2009AAS...21342506M&link_type=ABSTRACT

Reddick, R. M., Wechsler, R. H., Tinker, J. L., Behroozi, P. S., Jul. 2012. The Connection between Galaxies and Dark Matter Structures in the Local Universe. arXiv.org astro-ph.CO.
URL http://arxiv.org/abs/1207.2160

Schlegel, D., Abdalla, F., Abraham, T., Ahn, C., Prieto, C. A., Annis, J., Aubourg, E., Azzaro, M., Baltay, S. B. C., Baugh, C., Bebek, C., Becerril, S., Blanton, M., Bolton, A., Bromley, B., Cahn, R., Carton, P. H., Cervantes-Cota, J. L., Chu, Y., Cortes, M., Dawson, K., Dey, A., Dickinson, M., Diehl, H. T., Doel, P., Ealet, A., Edelstein, J., Eppelle, D., Escoffier, S., Evrard, A., Faccioli, L., Frenk, C., Geha, M., Gerdes, D., Gondolo, P., Gonzalez-Arroyo, A., Grossan, B., Heckman, T., Heetderks, H., Ho, S., Honscheid, K., Huterer, D., Ilbert, O., Ivans, I., Jelinsky, P., Jing, Y., Joyce, D., Kennedy, R., Kent, S., Kieda, D., Kim, A., Kim, C., Kneib, J. P., Kong, X., Kosowsky, A., Krishnan, K., Lahav, O., Lampton, M., LeBohec, S., Le Brun, V., Levi, M., Li, C., Liang, M., Lim, H., Lin, W., Linder, E., Lorenzon, W., de la Macorra, A., Magneville, C., Malina, R., Marinoni, C., Martinez, V., Majewski, S., Matheson, T., McCloskey, R., McDonald, P., McKay, T., McMahon, J., Menard, B., Miralda-Escude, J., Modjaz, M., Montero-Dorta, A., Morales, I., Mostek, N., Newman, J., Nichol, R., Nugent, P., Olsen, K., Padmanabhan, N., Palanque-Delabrouille, N., Park, I., Peacock, J., Percival, W., Perlmutter, S., Peroux, C., Petitjean, P., Prada, F., Prieto, E., Prochaska, J., Reil, K., Rockosi, C., Roe, N., Rollinde, E., Roodman, A., Ross, N., Rudnick, G., Ruhlmann-Kleider, V., Sanchez, J., Sawyer, D., Schimd, C., Schubnell, M., Scoccimaro, R., Seljak, U., Seo, H., Sheldon, E., Sholl, M., Shulte-Ladbeck, R., Slosar, A., Smith, D. S., Smoot, G., Springer, W., Stril, A., Szalay, A. S., Tao, C., Tarle, G., Taylor, E., Tilquin, A., Tinker, J., Valdes, F., Wang, J., Wang, T., Weaver, B. A., Weinberg, D., White, M., Wood-Vasey, M., Yang, J., Yeche, X. Y. C., Zakamska, N., Zentner, A., Zhai, C., Zhang, P., Jun. 2011. The BigBOSS Experiment. arXiv.org astro-ph.IM.
URL http://arxiv.org/abs/1106.1706v1

Tegmark, M., Taylor, A. N., Heavens, A. F., May 1997. Karhunen-Loeve Eigenvalue Problems in Cosmology: How Should We Tackle Large Data Sets? Astrophysical Journal v.480 480, 22.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1997ApJ...480...22T&link_type=ABSTRACT

Vives, S., Le Mignant, D., Madec, F., Jaquet, M., Prieto, E., Martin, L., Le Fèvre, O., Gunn, J., Carr, M., Smee, S., Barkhouser, R., Sugai, H., Tamura, N., Oct. 2012. A spectrograph instrument concept for the Prime Focus Spectrograph (PFS) on Subaru Telescope. arXiv.org astro-ph.IM.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=2012SPIE.8446E..4PV&link_type=ABSTRACT

Wechsler et al., 2013. Addgals i. arXiv.org.

White, R. A., Fink, R., Pisarski, R., Mar. 1991. FITS Formats for Space Data: ROSAT. Bulletin of the American Astronomical Society 23, 907.
URL http://adsabs.harvard.edu/cgi-bin/nph-data_query?bibcode=1991BAAS...23..907W&link_type=GIF

Appendix A. Functions within the Spectroscopic Pipeline

In this appendix, we discuss the basic functions of each module in the spectroscopic observation and analysis pipeline.

Appendix A.1. Target Selection

[need settle on details of our algorithm]

- **Inputs:** magnitudes $m \in \{g, r, i, z, y\}$; photometric redshifts $z_{\rm photo}$; positions; and unique galaxy

index; prescribed magnitude limits and color limits.

- **Output:** binary selection flags denoting whether a galaxy is selected for spectroscopic observation.

Appendix A.2. Survey Strategy

This function provides a complete Survey run simulator, incorporating realistic environmental and sky conditions. In this simulation, we aim to allow: 1) Field coverage optimization and a survey mask; 2) Target allocation and observation night efficiency ; and 3) Input for simulation pipeline and Figure of Merit analysis.

The module is divided into two main functions—the planner and the scheduler. The first one is in charge of allocating all the pointings necessary to cover one or more fields in the sky following a predefined hexagonal tiling pattern. The result is a list of tiles pending to be scheduled. This output mainly depends on the specified fields and the number of passes.

The scheduler then identifies the visibility of the fields during the year (under visibility only mode) and schedules the remaining pointings based on a set of observing nights. There are other relevant parameters such as airmass limit, first- and second-order moment seeing, exposure time, as well as moon brightness. At the end of the process, the module outputs a list of scheduled targets for each night with its observation time, airmass, seeing, sky background brightness and celestial coordinates.

This information is then passed to the fiber allocation module that, together with the position of the galaxies, positions the fibers in each of the scheduled tiles.

- **Inputs:**
- **Output:**

Appendix A.3. Fiber Allocation

The fiber allocation module takes as inputs the the center positions for each tile produced by the Survey Strategy module, the positions for the target galaxies produced by Target Selection and all the numbers describing the fibers: patrol radius, fiber diameter and number of fibers along the diameter of the hexagonal tile and the hexagon radius in degrees.

As a first step the fibers receive positions over the hexagonal tile following a padding where each fiber is surrounded by six equidistant fibers producing an hexagonal pattern. In this geometrical configuration each target in the sky can be reached at least (most) by three (four) fibers. In this process each fibers is assigned a unique ID.

The central part of the fiber allocation module is the algorithm that decides which galaxies are going to be matched by a fiber. We use two allocation algorithms.

The first one gives priority based on the local galaxy density. The motivation for this algorithm is give priority to galaxies in crowded regions regions. The first step in the algorithm is estimating for each galaxy the number of galaxies to be allocated within a patrol radius, n_p. We calculated for each spine a list of galaxies that can be reached, this list is ranked in descending order by n_p. For each spine the galaxy with the highest n_p is allocated. Then we check for fiber collisions: in the case of a collision, the two fibers are reset and the process iterates until the number of fiber collisions cannot be decreased or the number of collisions is zero.

The code has been implemented in Python. It takes about 25 seconds to run on a 3.8GHz processor for a single field of 8000 targets, observed twice with 4000 spines. Most of the run time is spent in re-setting the fiber collisions.

- **Inputs:** galaxy positions; fiber positions
- **Output:** flag on galaxies, denoting which fiber was matched (or -1 if no fiber is matched).

Appendix A.4. Throughput

In the *throughput* module, two kinds of efficiency as a function of wavelength are calculated—one for the physical optics in the main barrel of the instrument, as well as the spectrograph and fiber optics, and the other for the fiber optic aperture.

For the physical optics of the telescope, we include all major elements along the light path: from the top end footprint to the primary mirror and wide-field corrector. We add to this estimates of the efficiency

for the fiber positioning system **REF(i.e., Mohawk)** and the fiber efficiencies. For the dispersive element, we assume a volume-phase Holographic (VPH) grating **REF(need ref)**, whose efficiency spectrum is estimated via a VPH gsolver **REF(need ref)**. Next, for the spectrograph, we account for the collimators, the camera and the CCD.

Finally, we return the aperture calculation which has potential to highly influence the amount of galaxy light reaching the CCD. We convolve a Moffat (Beta), Gaussian, deVaucoleur and exponential profile to capture a variety of galaxy types and the degradation fo spot sizes; we calculate aperture losses for circular apertures, including offsets from the image center. **[Will: need details/explanations]**

While atmospheric power is used in our calculation, it is implemented in another module of the pipeline (Appendix A.5) and will be discussed there.

- **Inputs:** telescope physical optics efficiency spectra
- **Output:** throughput through full light path within telescope and instrument.

Appendix A.5. Reconstruction of Pure Galaxy Spectra and Noise Spectrum

Galaxy rest-frame spectral energy distributions are reconstructed using *kcorrect* **REF(Blanton 2003)** from a small set of template spectra, which are themselves derived via the non-negative matrix factorization technique **REF(BR 2007)**.

The output from the pure spectral reconstruction is a wavelength-redshifted and flux-dimmed SED for each galaxy, where the wavelengths are taken in Angstrom, and the flux is in units of $ergs\,cm^{-2}s^{-1}A^{-1}$. These are used in the noise generator to produce noise that includes the atmosphere, Poisson noise from counting photons and read noise from the spectrograph. Atmospheric absorption comes from the Palomar sky extinction model (from B. Oke and J. Gunn), and the atmospheric emission from optical sky background models from Gemini [2].

Details of the reconstruction and noise generation can be found in Appendix A2 of Cunha et al. (2012).

[2]Sky spectrum obtained from `http://www.gemini.edu/sciops/ObsProcess/obsConstraints/atm-models/skybg\_50\_10.dat`

- **Inputs:**
- **Output:**

Appendix A.6. Redshift Measurement

To measure galaxy redshifts from their spectra, we employ a cross-correlation method **REF(ref for this method)**, which compares a galaxy spectrum, s_i, with unknown redshift against a set of template spectra t with known redshifts.

On both the galaxy and some template spectra, we first perform a rebinning to a logarithmic wavelength grid **[WHY]**. This is followed by continuum subtraction using a moving average. Next, from the templates, we create a set of *eigen*templates, e, via principal component analysis. The eigentemplates are then incrementally shifted (on the logarithmic grid, corresponding to a redshift in the wavelength) until the average overlap with each continuum-subtracted galaxy spectrum is maximal, where the corresponding shift yields the spectroscopic redshift $z_{\rm spec} = \frac{\lambda^*}{\lambda_{\rm min}} - 1$. When tested on a truth catalog, the relative error of the spectroscopic redshift by this method is typically less than 0.5%.

- **Inputs:** galaxy spectra with noise, $s(\lambda)$; a set of template spectra $t(\lambda)$; wavelength range ($\lambda_{\rm min}$, $\lambda_{\rm max}$); number of bins N **REF(of what?)**
- **Output:** spectroscopic redshift $z_{\rm spec}$

Appendix A.7. Redshift Binning

The galaxies are binned in spectroscopic redshift and RA and DEc, according to a user-defined parameter for the number of bins. The result is the number density distribution $dn)/dzd\Omega$.

- **Inputs:** galaxy spectroscopic redshifts, galaxy positions, number of bins
- **Output:** number density distribution

Appendix A.8. Selection Mask

In general one would, **[XXX]**

In practice, for the current version of the pipeline, we merely need to measure the fraction of sky for which we have measured spectroscopic redshifts.

Appendix A.9. Estimation of Cosmological Parameters

The last step of the pipeline consists of forecasting the constraining power of a given survey configuration by analyzing the simulated galaxy catalogue.

From the redshifts and angular positions of the galaxies in the simulation we calculate the theoretically expected values for a number of correlation functions whose values depend on the cosmological parameters. In this paper, we choose to consider the angular matter power spectrum (i.e. two-point correlation function), because it can readily be compared to experimental data without assuming a cosmological model. The information contained in the angular power spectrum of a number of different redshift bins, as well as that between redshift bins, enables us to constrain the cosmological parameters by comparing theory with observation.

The constraints on the cosmological parameters expected for a given survey configuration and cosmological observable can be estimated using the Fisher matrix formalism before having conducted the actual experiment (Tegmark et al., 1997). Assuming that both the data and the parameters are Gaussian distributed, the Fisher matrix propagates the expected measurement errors down to uncertainties on the cosmological parameters derived from the experiment, assuming fiducial values. In the case of a tomographically analysed galaxy redshift survey, we can write the spherical harmonic decomposition of the observed angular power spectrum between two redshift bins x_i and x_j following (Hu and Jain, 2004) as

$$\tilde{C}_l^{x_i x_j} = C_l^{x_i x_j} + N_l^{x_i x_j} \tag{A.1}$$

where $N_l^{x_i x_j}$ is the contribution due to shot noise and given by

$$N_l^{x_i x_j} = \delta_{ij} \frac{1}{n_i} \tag{A.2}$$

The variable n_i denotes the angular galaxy density in bin i. Using these definitions the Fisher matrix of a given experiment with sky coverage $f_{\rm sky}$ can be computed with (Hu and Jain, 2004)

$$F_{\alpha\beta} = f_{\rm sky} \sum_l \frac{(2l+1)\,\Delta l}{2} \mathrm{Tr}\left[\mathbf{D}_{l\alpha} \tilde{\mathbf{C}}_l^{-1} \mathbf{D}_{l\beta} \tilde{\mathbf{C}}_l^{-1}\right] \tag{A.3}$$

where $\Delta l = 1$, $\tilde{\mathbf{C}}_l$ denotes the data covariance matrix and $\mathbf{D}_{l\alpha}$ contains the dependence of the observables on the cosmological parameters θ_α and whose elements are

$$[\mathbf{D}_{l\alpha}]^{ij} = \frac{\partial C_l^{x_i x_j}}{\partial \theta_\alpha}. \tag{A.4}$$

The Fisher matrix encodes the minimal standard deviation of each parameter around its maximum likelihood estimate (see e.g. (Kendall and Stuart, 1973)). This uncertainty is given by

$$\Delta\theta_\alpha \geq \sqrt{F^{-1}_{\alpha\alpha}} \tag{A.5}$$

To calculate the Fisher matrix we need to specify the parameter set to constrain and the fiducial values for these parameters. In our calculations we choose the set consisting of the seven cosmological parameters and fiducial values $\boldsymbol{\theta} = \{h = 0.7, \Omega_m = 0.3, \Omega_\Lambda = 0.7, w_0 = -0.95, w_a = 0.0, n_s = 1.0, \delta_H = 1843785.96\}$. The two parameters w_0 and w_a encode the time-dependence of the dark energy equation-of-state parameter which we choose to model as proposed by (Linder, 2003): $w(a) = w_0 + w_a(1-a)$.

Our Fisher matrix calculations are performed applying the Limber approximation (Limber, 1953) for the angular matter power spectrum :

$$C_{\rm l,\,Limber} = \frac{1}{c} \int H(z) \frac{W_i(z) W_j(z)}{\chi(z)^2} P\left(k = \frac{l+\frac{1}{2}}{\chi(z)}\right) dz. \tag{A.6}$$

The relevant cosmological quantities are calculated using the PyCosmo software **REF(Author et al.?)**. The redshift selection functions $W(z)$, fractional sky coverage $f_{\rm sky}$ and angular galaxy densities n_i are taken from the simulated and processed galaxy catalogue.

A better approach for future calculations is to use the exact expression for the angular power spectrum (not using the Limber approximation). Furthermore, the cosmological parameter module should be extended to compute parameter constraints for any possible observable in a galaxy survey.

- **Inputs:** number density distribution of galaxies $\mathrm{d}n/\mathrm{d}z\mathrm{d}\Omega$; sky fraction, $f_{\rm sky}$.

- **Output:** estimates of cosmological parameters, $\hat{\theta}$

Appendix B. Simulation Data

For this study we have used the mock galaxy catalogs created for the Dark Energy Survey based on the algorithm Adding Density Determined GAlaxies to Lightcone Simulations (ADDGALS; Busha et al., 2013; Wechsler et al., 2013). This algorithm attaches synthetic galaxies, including multiband photometry, to dark matter particles in a lightcone output from a dark matter N-body simulation and is designed to match the luminosities, colors, and clustering properties of galaxies. The catalog used here was based on a single "Carmen" simulation run as part of the LasDamas of simulations (McBride et al., 2009) [3]. This simulation modeled a flat ΛCDM universe with $\Omega_m = 0.25$ and $\sigma_8 = 0.8$ in a 1 Gpc/h box with 1120^3 particles. A 220 sq deg light cone extending out to z = 1.33 was created by pasting together 40 snapshot outputs.

The galaxy distribution for this mock catalog was created by first using an input luminosity function to generate a list of galaxies, and then adding the galaxies to the dark matter simulation using an empirically measured relationship between a galaxies magnitude, redshift, and local dark matter density, $P(\delta_{dm}|M_r, z)$ – the probability that a galaxy with magnitude M_r and redshift z resides in a region with local density δ_{dm}. This relation was tuned using a high resolution simulation combined with the SubHalo Abundance Matching technique that has been shown to reproduce the observed galaxy 2-point function to high accuracy (Conroy et al., 2007; KRAVTSOV et al., 2003; Reddick et al., 2012).

For the galaxy assignment algorithm, we choose a luminosity function that is similar to the SDSS luminosity function as measured in (Blanton et al., 2003), but evolves in such a way as to reproduce the higher redshift observations (e.g., SDSS-Stripe 82, AGES, GAMA, NDWFS and DEEP2). In particular, ϕ_* and M are varied as a function of redshift in accordance with the recent results from GAMA (Loveday et al., 2012).

[3] Further details regarding the simulations can be found at http://lss.phy.vanderbilt.edu/lasdamas/simulations.html

Once the galaxy positions have been assigned, photometric properties are added. Here, we use a training set of spectroscopic galaxies taken from SDSS DR5. For each galaxy, in both the training set and simulation, we measure Δ_5, the distance to the fifth (5th) nearest galaxy on the sky in a redshift bin. Each simulated galaxy is then assigned an SED based on drawing a random training-set galaxy with the appropriate magnitude and local density, k-correcting to the appropriate redshift, and projecting onto the desired filters. When doing the color assignment, the likelihood of assigning a red or a blue galaxy is smoothly varied as a function of redshift in order simultaneously reproduce the observed red fraction at low and high redshifts as observed in SDSS and DEEP2.

Appendix C. Computing Environment Requirements

The pipeline is designed to run on any system that has the required software.